ENHANCING STRUCTURE DIAGRAM GENERATION

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of United States Provisional Application Serial No. 60/119,654 entitled STRUCTURE DIAGRAM GENERATION filed on February 11, 1999, incorporated herein.

REFERENCE TO SOURCE CODE APPENDIX

A Microfiche ~~source code~~ appendix forms part of this application. The appendix, which includes a source code listing relating to an embodiment of the invention, includes 33 frames on 1 sheet of microfiche ~~30 pages.~~

This patent document (including the source code appendix) contains material that is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document as it appears in the Patent and Trademark Office file or records, but otherwise reserves all copyright rights whatsoever.

Background of the Invention

This application relates to enhancing structure diagram generation.

A molecule is typically represented in a computer by a connection table that identifies atoms in the molecule and specifies connections ("bonds") among the identified atoms. The connection table may also describe associated

properties such as atom type, bond order, charge, and stereochemistry. A diagrammatic representation of the molecule may be derived from the connection table. Examples of a connection table and a corresponding diagram are illustrated in Figs. 1A-1B (for clarity, hydrogen atoms are not shown).

In chemistry, with reference to Fig. 2, a chain of atoms that closes on itself is known as a ring. In the context of a ring or a ring system (see below), a bridge is a chain of atoms that begins at an origin point (which is an atom) in the ring or system, and connects back to the same ring or system at least two atoms away from the origin point, to form an additional ring. A chain that reconnects at the same origin point instead is known as "spiro". A chain that reconnects to an atom that is adjacent to the origin point is known as "fused".

A ring system, which is also known as a "cyclic system", is a group of rings such that (1) each ring shares one or more bonds with another ring in the group and (2) the group cannot be divided into smaller cyclic systems. An arrangement in which two rings are connected by a linking, non-cyclic ("acyclic") bond is considered to include two cyclic systems, not one. As used herein, "ring system" has a meaning consistent with an understanding that a spiro ring includes two distinct ring systems.

Summary of the Invention

A method and a system are provided for enhancing structure diagram generation ("SDG"). In SDG, aesthetic two-dimensional ("2-D") coordinates for use in a diagrammatic representation ("diagram") of a molecule are derived from a connection table for the molecule. SDG may also improve the aesthetic qualities of a chemical structure diagram having existing coordinates, if available. SDG is enhanced by expressing the symmetry present in the molecule, by making use of symmetry in the 2-D dynamics used to lay out rings and chains, by construction of bridges using an open polygon method together with a potential function, and by an elegant approach to the relative positioning of molecules ("free rectangle method").

Other features and advantages will become apparent from the following description, including the drawings, and from the claims.

Brief Description of the Drawings

Fig. 1A is an illustration of computer data.

Figs. 1B-1C, 3-4, 6-13 are illustrations of output produced by software.

Fig. 2 is an illustration of chemical structures.

Figs. 5, 14-16 are flow diagrams of computer-based procedures.

Detailed Description

Structure diagram generation ("SDG") is a process in which two dimensional ("2-D") coordinates are derived from a connection table for a structure, allowing a diagram of the molecule to be displayed or printed. SDG is described in detail in H. E. Helson, "Structure Diagram Generation", in "Reviews in Computational Chemistry", K. B. Lipkowitz and D. B. Boyd, Eds., Wiley-VCH, New York, 1999, Vol. 13, at 313-398, which is incorporated herein. This application is filed simultaneously with a United States patent application entitled DERIVING CHEMICAL STRUCTURAL INFORMATION, serial no. 09/502,810 filed February 11, 2000, which is incorporated herein.

The coordinates may be derived with or without preexisting coordinates. Cases without preexisting coordinates ("de novo" cases) are common and include chemical name translation, isomer enumeration, translation from a linear notation such as SMILES, nickname/superatom expansion, and automated structure elucidation.

In cases in which preexisting coordinates are available ("structure cleanup" cases), it may be possible to improve a structure diagram while preserving some or all existing stylistic choices. For example, if a structure diagram is drawn with or imported into a structure drawing program, the

program may be directed to "clean up" undesirable aspects of the structure diagram. In another example, diagram improvements may be needed in the case of a synthesis planning program, in which structure diagrams are generally well drawn but may have had bonds broken and reformed in awkward locations.

SDG may also be needed in conversions of structure diagrams from three-dimensional ("3-D") to 2-D. In at least some cases, structure diagrams that are stored and manipulated in a 3-D form may be converted to 2-D diagrams upon display to make the structure diagrams more easily recognizable to human users.

As a result, a connection table to which SDG is applied may have 2-D or 3-D coordinates or may lack coordinates.

SDG includes at least four possible stages ("phases"): perception, pre-assembly analysis, assembly, and post-assembly. The pre-assembly phase, if applicable, may include deriving a feature such as the shape of a ring system that is subsequently attached whole to an acyclic portion in the assembly phase. In the assembly phase, the neighbors of an atom that has been positioned (a "seed" atom) are each examined in turn, and are positioned at respective aesthetic angles and distances from the seed atom. Figs. 1A-1C illustrate an example. A connection table of a simple molecule (Fig. 1B) is shown in Fig. 1A.

In a specific embodiment, one of the atoms is arbitrarily chosen as a first seed atom. The neighbors of the first seed atom are positioned; each of the neighbors in turn takes the role of the seed atom in subsequent iterations, until all of the atoms are positioned, as shown in Fig. 1C for the connection table of Fig. 1A.

SDG is enhanced as described below.

In a first aspect of the enhancement, symmetry is used in the assembly phase, i.e., for general layout. Chemical structure diagrams that express molecular symmetry facilitate human interpretation of the chemical structures that are represented. For example, the presence of symmetry provides clues for the molecular substance's synthesis. Symmetry affects the substance's physical properties (particularly those affected by entropy), such as melting and boiling points, and heat of vaporization. Symmetry can affect the substance's light-bending properties. In particular, a substance that has a plane of symmetry is not "optically active". In general, since the human eye tends to recognize symmetry quickly, a diagram that expresses a molecule's symmetry allows the symmetrical characteristics of the molecule to be rapidly perceived by a human viewer.

According to the enhancement, when a diagram is to be produced for a molecule, symmetry inherent in the molecule is perceived, and during layout of

the structure diagram, representations of atoms and bonds are positioned to express the perceived symmetry. In a specific implementation, a plane of symmetry (also known as a mirror plane) perceived in a molecule is expressed vertically or horizontally (see Fig. 3).

In a first step in using symmetry in general layout (see Figs. 4 and 14), an instance of symmetry is determined (step 1010). Detection of symmetry is described in M. Razinger, K. Balasubramanian, and M. E. Munk, "Graph Automorphism Perception Algorithms in Computer-Enhanced Structure Elucidation", *J. Chem. Inf. Comput. Sci.*, **33**, 197 (1993). The instance of symmetry may be based on one or more of rotation, reflection (see b. in Fig. 4), inversion, and translation, and may or may not take stereochemistry into account. A particularly effective combination in the determination is an instance of symmetry based on rotation and reflection, and a flexible incorporation of stereochemistry. If a full consideration of stereochemistry does not reveal an instance of symmetry, a partial consideration of stereochemistry, e.g., of double bond stereochemistry only, is employed. The instance of symmetry is here represented as a list of orbits, i.e., a list of groups of equivalent atoms and bonds.

Additionally, a "pivot" point is determined for each orbit (step 1020). The pivot point is determined to be the one or more atoms or bonds that resides at the graph-theoretic center of the atoms and bonds in the orbit, i.e., those atoms (bonds) having the smallest value of the largest graph-theoretic distance to any other atom (bond) in the orbit. The graph-theoretic distance between two atoms (bonds) is equal to the number of bonds (atoms) in the shortest path between them. For example, in Figs. 3-4, the nitrogen atom is determined to be the pivot point; in n-butane, the central bond is determined to be the pivot point, and in 1,2,3-trimethylcyclopropane, all cyclic atoms and bonds are determined to be the pivot point.

The "order" of each orbit for each instance of symmetry is also determined (step 1030). The order indicates whether the instance corresponds to a two-fold rotation, a three-fold rotation, a four-fold rotation, and so on, or a reflection. In cases in which the symmetry of an orbit includes both reflection and an N-fold rotation, N being greater than 2, it is advantageous to treat the instance as having an order indicating that the instance corresponds to the N-fold rotation. Thus, rotational symmetry takes priority over reflection if the associated rotation is at least three-fold.

When an atom or bond is positioned during the assembly phase (step 1040) (see Figs. 1, 4), attention is paid to whether the atom or bond belongs to one of the determined instances of symmetry. In a case in which the atom or bond so belongs, after the atom or bond is positioned, other atoms or bonds, respectively, that belong to the same instance are positioned immediately thereafter (step 1050) (see Figs. 4-5). If the type of symmetry involved is reflection (see Fig. 5), the other atom or bond is placed on the opposite side of the mirror line that runs through the pivot point of the group in the instance. In such cases, the direction may be arbitrary if only two atoms have been placed. If the type of symmetry involved is rotation, the symmetrically equivalent atoms or bonds are positioned at appropriate rotational points, based on the pivot point. First positioning an atom or bond that represents the pivot point facilitates symmetric positioning but is not always possible, such as when multiple regions of independent symmetry are involved (e.g., in an unsymmetrical ether, each end of which is locally symmetric).

After all atoms and bonds have been placed, the structure diagram is rotated so that its mirror plane is horizontal or vertical (step 1060) (see Fig. 3).

In another aspect of the enhancement of SDG, symmetry is used in a "dynamics" method of layout. A 2-D version of molecular dynamics is used in

some situations to lay out structure diagrams of molecules in connection with designing new ring systems, improving existing ring systems, or laying out or improving acyclic portions. Such an effort may use a predefined set of optimal bond lengths and angles ("parameters"), or may seek to equalize adjacent lengths and angles. The process is iterative, wherein in each iteration the difference between a current parameter and an optimal parameter is calculated for each atom and bond, and is interpreted as a corrective force on the atom or bond, which affects the position of the atom or bond as submitted to the next iteration. The iterative process continues until the net corrective force on every atom or bond is zero or nearly zero, so that the structure diagram for the molecule is determined to be at equilibrium.

A method of adding symmetry as a parameter in dynamic ring layout is now described (Fig. 15). The concepts presented are also applicable to acyclic systems. Dynamic ring layout in general is described in H. E. Helson, Ph.D., Thesis, "Simulation of Carbene Chemistry and Other Problems in Computer-Assisted Organic Synthesis.", Purdue University, 1993; H. E. Helson and W. L. Jorgensen, J. Chem. Inf. Comput. Sci., 34, 962 (1994), "Computer-Assisted Mechanistic Evaluation of Organic Reactions. 25. Structure Diagram Positioning"; and H. E. Helson, "Structure Diagram Generation", in "Reviews in

Computational Chemistry", K. B. Lipkowitz and D. B. Boyd, Eds., Wiley-VCH, New York, 1999, Vol. 13, at 313-398. As shown below, a symmetry term is incorporated in a force field, which drives symmetrically equivalent regions in different parts of the structure diagram toward a common appearance. Instances of symmetry are determined (step 2010). The symmetry detection referenced above is an example of such a determination. In a specific implementation, the instances are determined based on rotation and reflection, without regard for bond orders or types, atom characteristics (e.g., mass, type, charge), or acyclic portions, and the determination focuses exclusively on the locations of the bonds. The instances of symmetry may be determined without supplied coordinates. The determination takes double bond isomerism into account: E and Z isomers are recognized as not being equivalent. Specific implementations may also take into account 2-D graphics-based characteristics not normally connected with molecular symmetry, such as bond zig-zags, or whether a bond is "exterior" or "interior", i.e., whether or not the bond has a clear path to the edge of the drawing area.

The instance of symmetry, regardless of character and origin, may be represented in any of several ways. In a specific implementation, the instance is represented by two lists of groups: a list of equivalent triplets of atoms, and a

list of equivalent pairs of bonds (see Fig. 6, in which the top and bottom sequences illustrate equivalent bonds and atom triplets, respectively, and each dot marks the central atom in a triplet).

In each iteration for each triplet, a respective force term ("F_a") is added for the atom in the center of the triplet (step 2020). An optimal interior angle ("optimal angle") of the triplet of atoms is derived, as the average of the interior angles of all the triplets in an orbit, i.e., in a group of symmetrically equivalent atoms or bonds. F_a is based on, and in a specific implementation is proportional to, the difference between the optimal angle and the current angle. F_a acts along the angle's bisector, in a direction that would bring the angle closer to the optimal angle. F_a may compete with other terms, such as a bond angle term for equalizing adjacent bond angles.

In each iteration, another respective force term ("F_b") is added for each symmetric bond (step 2030). F_b has the effect of lengthening or shortening a bond to make the bond's length more similar to the lengths of the other bonds in the orbit. A bond's length is changed by moving the atoms at the bond's endpoints closer together or farther apart. Thus F_b is expressed by treating F_b as a force on each of its two adjacent atoms. F_b may compete with other terms, such as a bond length term for equalizing adjacent bond lengths.

During each iteration, a net force on each atom is calculated, as the sum of the forces including F_a and F_b acting on the atom (step 2040). The position of each atom is moved by an amount proportional to the respective net force. In a specific implementation, the iterative process is determined to be complete when the largest net force to be accounted for in the iteration is smaller than a specified threshold size (step 2050).

Fig. 7 illustrates an example of net forces and the iterative evolution. In Fig. 7, double arrows show the forces due to symmetry on selected atoms and bonds, single dashed arrows represent bond angle forces re-expressed as atom translation, and other forces not related to symmetry are omitted for clarity and to reduce clutter. The rightmost structure diagram in Fig. 7 represents an improvement over the leftmost structure diagram.

Construction of bridged cyclic systems may involve problems of atom and bond overlap, and irregular angles. In another aspect of the enhancement of SDG, bridges in cyclic systems are constructed using an open polygon method in conjunction with a potential function. In an example illustrated in Fig. 8, SDG has already produced two rings that are part of a tricyclic system. In a regular polygon method, a third ring (indicated by dashed lines in Fig. 8) is attached by constructing the third ring as a regular polygon, so as to fuse the third ring to

either one of the rings already produced. In such a case, as shown in the top two example sequences in Fig. 8, uneven coordinates are produced. Alternatively in the regular polygon method, the regular polygon can be attached directly at the two bridgehead atoms, as shown in the bottom example sequence in Fig. 8, but uneven results are still achieved. By contrast, the open polygon method is able to generate evenly spaced coordinates between the two termini where the ring will be attached, as shown in an example sequence in Fig. 9, in which a five-membered ring is fused onto a bicyclo[6.1.0] system. See H. E. Helson, "Structure Diagram Generation", in "Reviews in Computational Chemistry", K. B. Lipkowitz and D. B. Boyd, Eds., Wiley-VCH, New York, 1999, Vol. 13, at 313-398, which is incorporated herein.

In the open polygon method, coordinates of missing points are derived from two grounding points, the number of missing points, and an optimal bond length ("*d*"), such that, as shown in Fig. 9, interior angles ("β") at the two grounding points are equal and the remaining interior bond angles ("α") are all equal.

The open polygon method can be used to create bridges. In Fig. 9, the two grounding points correspond to the two bridgehead atoms. In at least some cases, however, the resulting bridge may be determined to be too close or too far

away from the base ring skeleton such that the resulting bridge crowds the base ring skeleton. As a result, the circumference of the bridge is varied by varying the value of *d* in Fig. 9.

Fig. 10 illustrates an example of an application of the open polygon method to bridge construction. The leftmost structure diagram represents the preexisting ring skeleton to which the bridge will be affixed. The other structure diagrams represent the results of applying the open polygon method using various values of *d*. The one producing the least congestion, among other factors, is chosen. More specifically, the value of *d* that is selected is the value that (1) produces the least congestion between the bridge and the base ring skeleton, as measured, for example, by a two-body inverse-distance squared potential function, (2) does not produce near-linear bond angles, i.e., does not produce an α that is nearly 180 degrees, and (3) uses a bond length *d* that is close to the optimal bond length, as may be expressed in a weighted function, such as:

$$\text{Rating} = c1 * \text{Congestion} + c_2 * \max(0, (180 - \alpha) - \text{threshold}) + c_3 * |\, scale - 1.0 \,|$$

In such a function, c_1, c_2 and c_3 are constants determined in a specific implementation; and *scale* is the ratio of d to the standard bond length. In a specific implementation, the bond angle term is active only above a certain threshold, such as 120 degrees. The version of the bridge that minimizes the rating is chosen.

Fig. 11 illustrates an application of the ratings method. Each line starting with "Rating for" indicates a rating computed for a particular bond length. For example, the second such line reports a rating of 313.185 for a bond length scale of 0.5 where the contribution for congestion is 45.185, the contribution for a non-unitary bond length is 40.00, and the contribution for a non-linear bond angle of 177 is 228. With respect to Fig. 11, the bond length scale that achieves the lowest rating and is therefore selected is 1.3.

In another aspect of the enhancement of SDG, a placement procedure is executed to arrange molecule structure diagrams closely together without overlapping. In at least some cases, the procedure is executed as a final step of SDG, after the molecule structure diagrams have been produced individually. The procedure is analytic in that the procedure does not rely on an indefinite number of iterations and is not affected by the starting positions of the components.

A specific implementation of the procedure is now described (Fig. 16), and an example as described below is illustrated in Fig. 13. A set of molecule structure diagrams and associated coordinates are acquired (step 3010). Each molecule structure diagram is represented by a conceptual box, defined by the structure diagram's smallest enclosing rectangle plus a small margin.

A "free rectangle" list is maintained that keeps track of which areas of the display area are unused (step 3020). The list is initialized to one free rectangle that occupies all of 2-D space and extends from negative infinity to positive infinity in both X and Y dimensions.

The boxes are sorted, and each is treated as follows, in order of decreasing area (step 3030). A free rectangle is selected that is closest to the center of the boxes and that is large enough to contain the instant box (step 3040). The center of a collection ("conglomeration") of boxes is defined as the average of the centers of the boxes weighted by the boxes' respective areas, or, as the center of the smallest rectangle that can enclose the boxes. The instant box is positioned flush with that corner of the free rectangle that is closest to the center of the growing collection (initially at coordinates (0,0)), and is imprinted on the free rectangle (step 3050). In imprinting, the original free rectangle is replaced by zero or more new free rectangles. New free rectangles may be created in the

leftover space, i.e., wherever the box does not occlude the original rectangle (see Fig. 12, which illustrates an example of the evolution of free rectangles and box placement). In an analogy in which a cookie cutter represents the box and dough represents the free rectangle, the dough underneath the cutter is discarded and the remaining areas of dough represent the leftover space that becomes the new free rectangles. In at least some cases, the sum of areas is not conserved, because each new free rectangle expands in both X and Y dimensions to the furthest extent possible without penetrating existing boxes. Several overlapping free rectangles may be necessary to fully span the leftover space (see Fig. 12). A free rectangle that could not contain the smallest box is not created.

In a specific implementation, overlapping free rectangles may be merged to help avoid a profusion of inconsequential free rectangles (step 3060). For example, rules may be enforced that dictate that two free rectangles should be merged such that the resulting free rectangle does not extend over any points not contained in either progenitor, provided that the percentage of area lost in the merger is less than a specified size, such as ten percent of the original area.

The conglomerate of boxes is translated so that its center is at coordinates (0,0) (step 3070). The molecule diagram coordinates are translated so that their centers coincide with their corresponding box centers (step 3080).

A practical example of the molecule arrangement procedure is illustrated in Fig. 13. Initially, a collection of molecules is presented to be positioned. Corresponding enclosing boxes are identified in an order of decreasing area (a. to d.), and are positioned one by one in the same order (i.e., a. first, d. last), to produce a space-efficient, non-overlapping arrangement as shown.

All or a portion of the procedures described above may be implemented in hardware or software, or a combination of both. In at least some cases, it is advantageous if the technique is implemented in computer programs executing on one or more programmable computers, such as a personal computer running or able to run an operating system such as UNIX, Linux, Microsoft Windows 95, 98, 2000, or NT, or MacOS, that each include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device such as a keyboard, and at least one output device. Program code is applied to data entered using the input device to perform the technique described above and to generate output information.

The output information is applied to one or more output devices such as a display screen of the computer.

In at least some cases, it is advantageous if each program is implemented in a high level procedural or object-oriented programming language such as Perl, C, C++, or Java to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

In at least some cases, it is advantageous if each such computer program is stored on a storage medium or device, such as ROM or optical or magnetic disc, that is readable by a general or special purpose programmable computer for configuring and operating the computer when the storage medium or device is read by the computer to perform the procedures described in this document. The system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner.

Other embodiments are within the scope of the following claims. For example, a non-human entity such as a computer program may serve as a source for input information such as the connection table or as a recipient of output

information such as diagrammatic data. In another example, one or more techniques based on the description herein may be applied to adapting structure diagrams for purposes other than presentation to a human user.